Large Bayesian Vector Autoregressions with Stochastic Volatility and Non-Conjugate Priors*

Andrea Carriero
Queen Mary, University of London
a.carriero@qmul.ac.uk

Todd E. Clark
Federal Reserve Bank of Cleveland
todd.clark@clev.frb.org

Massimiliano Marcellino
Bocconi University, IGIER and CEPR
massimiliano.marcellino@unibocconi.it

This draft: April 2018

Abstract

Recent research has shown that a reliable vector autoregression (VAR) for forecasting and structural analysis of macroeconomic data requires a large set of variables and modeling time variation in their volatilities. Yet, there are no papers that provide a general solution for combining these features, due to computational complexity. Moreover, homoskedastic Bayesian VARs for large datasets so far restrict substantially the allowed prior distributions on the parameters. In this paper we propose a new Bayesian estimation procedure for (possibly very large) VARs featuring time-varying volatilities and general priors. We show that indeed empirically the new estimation procedure performs well in applications to both structural analysis and out-of-sample forecasting.

J.E.L. Classification: C11, C13, C33, C53.

*We would like to thank the Editor Frank Diebold, two anonymous referees, Joshua Chan, Ana Galvao, Gary Koop, Dimitris Korobilis, Haroon Mumtaz, Davide Pettenuzzo, Anna Simoni and participants at seminars and conferences at the Banque de France, ECB, Bank of England, and the University of Pennsylvania conference on Big Data in Predictive Dynamic Econometric Modeling for useful comments on a previous version. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Cleveland or the Federal Reserve System. Carriero gratefully acknowledges support for this work from the Economic and Social Research Council [ES/K010611/1].

1 Introduction

The recent literature has shown that two main ingredients are key for the specification of a good vector autoregression (VAR) for forecasting and structural analysis of macroeconomic data: a large cross section of macroeconomic variables, and time variation in their volatilities. Banbura, Giannone, and Reichlin (2010), Carriero, Clark, and Marcellino (2015), Giannone, Lenza, and Primiceri (2015), and Koop (2013) point out that larger systems perform better than smaller systems in forecasting and structural analysis. Clark (2011), Clark and Ravazzolo (2015), Cogley and Sargent (2005), D'Agostino, Gambetti and Giannone (2013), and Primiceri (2005) highlight the importance of time variation in the volatilities, most typically modeled as stochastic volatility.

Both of these ingredients are commonly accommodated with Bayesian estimation methods, which rely on a posterior distribution that is the product of the likelihood and the prior distribution. Bayesian shrinkage is helpful to accommodating the many parameters that come with large models, and Bayesian computation is helpful for stochastic volatility. Accordingly, in this paper we focus on Bayesian estimation methods for VARs with stochastic volatility. However, the basic computational problem we describe and our solution would also apply with other formulations of time-varying error volatility, such as GARCH.

Although the aforementioned literature suggests that it would be ideal to combine a large cross-section of variables with stochastic volatility in a VAR, to this point computational challenges have precluded doing so in a general way. To understand why, it is helpful to begin with the familiar VAR with constant error variances (conditional homoskedasticity). With N variables, p lags, and an intercept included, the model has $N(Np+1)$ regression coefficients. Homoskedastic VARs are SUR models featuring the same set of regressors in each equation, which is commonly referred to as symmetry across equations. The symmetry yields a Kronecker structure in the variance-covariance matrix of the VAR's coefficients and means that, in a maximum likelihood (ML) context, the model can be estimated via OLS equation by equation. In a Bayesian setting with conditionally homoskedastic errors, as long as the prior distribution governing the VAR's coefficients takes a Kronecker structure, the posterior distribution also has a Kronecker structure. That structure makes feasible computation with large (conditionally homoskedastic) models, by reducing key computations to manipulation of Kronecker products rather than larger matrices. In particular, existing work with large Bayesian VARs has relied on the natural conjugate Normal-Wishart prior proposed by Kadiyala and Karlsson (1993, 1997).[1] For example, Banbura, Giannone,

[1] See Geweke and Whiteman (2006) and Karlsson (2013) for excellent surveys on priors for Bayesian VARs. Studies including Chib and Greenberg (1995) and Korobilis and Pettenuzzo (2017) have developed alternative

and Reichlin (2010) estimate a homoskedastic Bayesian VAR with 130 variables, using the natural conjugate prior.

Adding the macro literature's common formulation of stochastic volatility to the VAR breaks the symmetry of the model and the Kronecker structure of the posterior distribution of the VAR's coefficients, effectively precluding the estimation of large models. To understand why, it helps to return to the ML context. With the error term of each equation of the VAR featuring time-varying volatility, dividing the terms of each equation by its time-varying variance yields a system of equations with homoskedastic errors but with different regressors in each equation. The system of equations lacks the symmetry that applies with conditional homoskedasticity. The ML estimator is obtained by GLS applied to the system of equations instead of OLS applied equation-by-equation. The speed of computation that comes with the Kronecker structure of the conditional homoskedasticity case is lost; estimation must apply to the joint system of equations and involve very large matrices in large models. In the Bayesian setting, computations with the posterior distribution of the VAR's coefficients involve a variance matrix with $N(Np+1)$ rows and columns and without a Kronecker structure. The size of this matrix increases with the square of the number of variables in the model, making CPU time requirements highly nonlinear in the number of variables. Estimation becomes rapidly unmanageable as the number of variables increases.

With large models, even with conditional homoskedasticity, the same challenge to estimation can arise if the prior distribution of the VAR's coefficients lacks the symmetry that comes with a Kronecker structure. In this case, despite the symmetry of the model's equations, the posterior distribution of the VAR's coefficients lacks a Kronecker structure, and posterior computations involve a variance matrix with $N(Np+1)$ rows and columns, the size of which increases with N^2. Although much work has made use of priors with symmetry to circumvent computational challenges, priors without symmetry can be useful. One such prior is the original Litterman (1986) implementation of the so-called Minnesota prior, which puts additional shrinkage on the lags of all the variables other than the dependent variable of the i-th VAR equation, with the idea that these lags should be less relevant than the lag of the dependent variable itself. In this case the prior is not symmetric across equations and therefore, despite symmetry of the model's equations, the resulting posterior lacks a Kronecker structure, which implies that the model must be estimated as a system of equations. Incidentally, it is for this reason that Litterman (1986) assumed a (fixed) diagonal prior variance for the disturbance term, since this assumption allows one to estimate his model equation by equation. In general, the common Normal-diffuse and independent

Bayesian approaches for large VARs (without stochastic volatility) that rely on hierarchical priors.

Normal-Wishart priors also introduce asymmetry in the posterior of the VAR coefficients.

In the face of the computational challenges to including stochastic volatility in large VARs, a few studies have developed approaches that make use of some shortcuts. Koop and Korobilis (2013) and Koop, Korobilis, and Pettenuzzo (2016) propose a computational (not fully Bayesian) shortcut that allows for time-varying volatility using, roughly speaking, a form of exponential smoothing of volatility that allows them to estimate a large VAR. However, the resulting estimates are not fully Bayesian and do not allow, for example, computing the uncertainty around the volatility estimates in a coherent fashion. Carriero, Clark, and Marcellino (2016) instead make fully Bayesian inference feasible by assuming a specific factor structure for the volatilities in the VAR. Although their evidence indicates that the proposed model improves over an homoskedastic VAR in density forecasting, the restrictions implied by the factor structure do not necessarily hold in a typical dataset of macroeconomic and financial variables, especially so as the cross-sectional dimension grows.

In this paper, we develop a new, more general Bayesian estimation procedure — one without shortcuts — for large VARs with time-varying volatility or asymmetric (non-conjugate) priors. Our procedure is based on a simple triangularization of the VAR, which permits sampling the posterior of the VAR's coefficients by drawing them equation by equation. With N variables in the model, this reduces the computational complexity to the order N^4, which is considerably faster than the complexity of N^6 arising from the traditional algorithms. Our new algorithm is very simple and, importantly, it can be easily inserted in any pre-existing algorithm for Bayesian estimation of VARs. With our method, estimation of very large VARs with stochastic volatility and non-conjugate priors becomes feasible, and this is important both for reduced form applications, such as forecasting or constructing coincident and leading indicators, and for structural applications, such as computing responses to structural shocks or variance decompositions. Hence, our method also paves the way for a large number of empirical applications.

As an example and illustration, we estimate a VAR with stochastic volatility (VAR-SV), using a cross-section of 125 variables for the U.S. extracted from the dataset in McCracken and Ng (2016). Results show substantial homogeneity in the estimated volatility patterns for variables belonging to the same group, such as industrial production and producer price components or interest rates at different maturities, but there is some heterogeneity across groups of variables. When we use this very large VAR-SV to analyze U.S. monetary policy shocks and their transmission, we obtain impulse responses similar to those from homoskedastic specifications such as Bernanke, Boivin and Eliasz (2005) and Banbura, Giannone and Reichlin (2010), although with our time-varying variances, the size of the policy shock and

overall response magnitudes are varying over time.

Finally, we analyze the effect that the size of the cross-section and the time variation in the volatilities has on out-of-sample forecasting performance. We compare small and medium-sized (20 variable) VARs for the U.S., with and without stochastic volatility, in a recursive out-of-sample exercise, where the inclusion of the medium-sized VAR-SV is only feasible thanks to our new estimation method. We show that, jointly, the inclusion of time-varying volatilities and the use of a large dataset improve point and density forecasts for macroeconomic variables, with gains that are larger than what would be obtained by using these two ingredients separately.

The paper is structured as follows. Sections 2 and 3 introduce the model and develop the estimation method. Section 4 illustrates the gains obtained by our algorithm in terms of computing time, convergence, and mixing. Section 5 discusses the empirical application. Section 6 presents the out-of-sample forecasting exercise. Section 7 concludes.

2 Challenges in estimating large VARs with asymmetric priors and time varying volatilities

2.1 The model

Consider the following VAR with stochastic volatility:

$$y_t = \Pi_0 + \Pi(L)y_{t-1} + v_t; \quad (1)$$

$$v_t = A^{-1}\Lambda_t^{0.5}\epsilon_t,\ \epsilon_t \sim iid\ N(0, I_N), \quad (2)$$

where $t = 1, ..., T$, the dimension of the vectors y_t, v_t and ϵ_t is N, $\Pi(L) = \Pi_1 L + \Pi_2 L^2 + \cdots + \Pi_p L^p$, Λ_t is a diagonal matrix with generic j-th element $\lambda_{j,t}$, and A^{-1} is a lower triangular matrix with ones on its main diagonal.

The specification above implies a time-varying variance matrix, Σ_t, for the disturbances v_t. Importantly, our approach can be applied to any model featuring a time-varying error variance matrix, regardless of how its time variation is modeled. In what follows we use stochastic volatility and a factorization of Σ_t common in many macroeconomic applications:

$$\Sigma_t \equiv Var(v_t) = A^{-1}\Lambda_t A^{-1\prime}. \quad (3)$$

The diagonality of the matrix Λ_t implies that the generic j-th element of the rescaled VAR disturbances $\tilde{v}_t = Av_t$ is $\tilde{v}_{j,t} = \lambda_{j,t}^{0.5}\epsilon_{jt}$. Taking logs of squares of $\tilde{v}_{j,t}$ yields the following set of observation equations:

$$\ln \tilde{v}_{j,t}^2 = \ln \lambda_{j,t} + \ln \epsilon_{j,t}^2,\ j = 1, \ldots, N. \quad (4)$$

The model is completed by specifying laws of motion for the unobserved states:

$$\ln \lambda_{j,t} = \ln \lambda_{j,t-1} + e_{j,t},\ j = 1, \ldots, N, \tag{5}$$

where the vector of innovations to volatilities e_t is *i.i.d.* $N(0, \Phi)$, with full variance matrix Φ as in Primiceri (2005), not diagonal as in Cogley and Sargent (2005).[2]

In equation (2) we do not allow the elements in A^{-1} to vary over time. We do so because Primiceri (2005) found little variation in such coefficients (as we did in robustness checks of Carriero, Clark, and Marcellino 2017 with larger models), and specifying variation in these coefficients would imply an additional $N(N-1)/2$ state equations. Note, however, that even if one were to specify A^{-1} as time-varying, this would not impact the main computational advantage arising from the estimation method we propose below, as the main bottleneck in estimating large VARs is the inversion of the variance matrix of the $\Pi(L)$ coefficients, not the simulation of the drifting covariances and volatilities. That said, although our proposed approach solves the main bottleneck due to the size of the variance matrix of the VAR coefficients, in large systems (e.g., 30 or more variables), the estimation of a time-varying A matrix would still be challenging. For example, with a 30 variable model, an additional 29 equations with 435 states in A would need to be estimated.[3] Finally, as a simpler or less computationally challenging matter, one can modify equation (5) so that the states $\ln \lambda_{j,t}$ follow an autoregressive process rather than a random walk, but again this is not essential to the main point we make in this paper.

In a Bayesian setting, to estimate the model the likelihood needs to be combined with a prior distribution for the unobserved states Λ_t and the model coefficients Π, A, and Φ, where $\Pi = (\Pi_0, \Pi_1, ..., \Pi_p)'$ is a $(Np+1) \times N$ matrix. Typically the priors for the coefficient

[2] The specification of Primiceri (2005) is more general and allows for the volatilities to be hit by a common shock (while their conditional means are modeled independently of one another). However, as N gets large with respect to T, allowing correlations across variables might become problematic. In the case of a full Φ matrix, innovations to the volatility are modeled with an inverse Wishart prior, which needs to use at least $N+2$ degrees of freedom to be proper. With large N, this makes the prior highly informative, more so with quarterly data than monthly. A researcher worried about that could treat the innovations as independent and draw them from individual inverse gamma distributions, as in Cogley and Sargent (2005). Of course this amounts to imposing the restriction that both the prior and the likelihood have a diagonal Φ matrix, which can be seen as an even more informative prior than the Wishart one.

[3] There is also a related problem on how to calibrate the prior on so many state variables under an approach like that of Primiceri (2005), since pre-sample data are very limited.

blocks of the model are specified as follows:

$$\begin{aligned} \text{vec}(\Pi) &\sim N(\text{vec}(\underline{\mu}_\Pi), \underline{\Omega}_\Pi); && (6) \\ A &\sim N(\underline{\mu}_A, \underline{\Omega}_A); && (7) \\ \Phi &\sim IW(\underline{d}_\Phi \cdot \underline{\Phi}, \underline{d}_\Phi), && (8) \end{aligned}$$

where $N(\mu, \Omega)$ denotes a multivariate Normal distribution with mean μ and variance Ω and $IW(\underline{d}_\Phi \cdot \underline{\Phi}, \underline{d}_\Phi)$ denotes an inverse Wishart distribution with $\underline{d}_\Phi$ degrees of freedom and scale matrix $\underline{d}_\Phi \cdot \underline{\Phi}$. The model is completed by eliciting a prior for the initial value of the state variables Λ_t, for which we use an uninformative Normal distribution.

2.2 Model estimation

The VAR-SV model is typically estimated as follows. First, the conditional posterior distributions of all the coefficients blocks are derived:

$$\begin{aligned} \text{vec}(\Pi)|A, \Lambda_T, y_T &\sim N(\text{vec}(\bar{\mu}_\Pi), \overline{\Omega}_\Pi); && (9) \\ A|\Pi, \Lambda_T, y_T &\sim N(\bar{\mu}_A, \overline{\Omega}_A); && (10) \\ \Phi|\Lambda_T, y_T &\sim IW((\underline{d}_\Phi + T) \cdot \bar{\Phi}, \underline{d}_\Phi + T), && (11) \end{aligned}$$

where Λ_T and y_T denote the history of the states and data up to time T, and where the posterior moments $\bar{\mu}_\Pi$, $\overline{\Omega}_\Pi$, $\bar{\mu}_A$, $\overline{\Omega}_A$, and $\bar{\Phi}$ can be derived by combining prior moments and likelihood moments.[4] In particular, defining $X_t = [1, y'_{t-1}, ..., y'_{t-p}]'$ as the $(Np+1)$-dimensional vector collecting the regressors in equation (1), the mean and variance of the conditional posterior of the VAR's coefficients are given by

$$\begin{aligned} \text{vec}(\bar{\mu}_\Pi) &= \bar{\Omega}_\Pi \left\{ \text{vec}\left(\sum_{t=1}^{T} X_t y'_t \Sigma_t^{-1} \right) + \underline{\Omega}_\Pi^{-1} \text{vec}(\underline{\mu}_\Pi) \right\}; && (12) \\ \overline{\Omega}_\Pi^{-1} &= \underline{\Omega}_\Pi^{-1} + \sum_{t=1}^{T} (\Sigma_t^{-1} \otimes X_t X'_t). && (13) \end{aligned}$$

Defining the collection of model coefficients as $\Theta = \{\Pi, A, \Phi\}$, a step of a Gibbs sampler cycling through (9)-(11) provides a draw from the joint posterior distribution $p(\Theta|\Lambda_T, y_T)$. Conditional on this draw, a draw from the distribution of the states $p(\Lambda_T|\Theta, y_T)$ is obtained using the observation and transition equations (4) and (5), by using a mixture of normals

[4] Note that knowledge of the full history of the states Λ_T renders redundant conditioning on the hyperparameters Φ regulating the law of motions of such states when drawing Π and A, as well as conditioning on Π and A when drawing Φ.

approximation and multi-move algorithm proposed by Kim, Shephard, and Chib (1998).[5] Cycling through $p(\Theta|\Lambda_T, y_T)$ and $p(\Lambda_T|\Theta, y_T)$ provides the joint posterior of the model coefficients and unobserved states $p(\Theta, \Lambda_T|y_T)$.

In this paper we are interested in one specific step of the algorithm: the draw from $\text{vec}(\Pi)|A, \Lambda_T, y_T$ described in equation (9). The main problem is that this step involves the manipulation of the variance matrix of the coefficients Π, which is a square matrix of dimension $N(Np+1)$.

Consider drawing $m = 1, ..., M$ draws from the posterior of Π. To perform a draw Π^m from (9), one needs to draw a $N(Np+1)-$dimensional random vector (distributed as a standard Gaussian), denoted rand, and to compute:

$$\text{vec}(\Pi^m) = \bar{\Omega}_\Pi \left\{ \text{vec}\left(\sum_{t=1}^{T} X_t y_t' \Sigma_t^{-1} \right) + \underline{\Omega}_\Pi^{-1} \text{vec}(\underline{\mu}_\Pi) \right\} + chol(\bar{\Omega}_\Pi) \times \text{rand}\,. \tag{14}$$

This calculation involves computations of the order of $4 \times O(N^6)$. Indeed, it is necessary to: i) compute the matrix $\bar{\Omega}_\Pi$ by the inversion given in equation (13); ii) compute its Cholesky factor $chol(\bar{\Omega}_\Pi)$; and iii) multiply the matrices obtained in i) and ii) by the vector in the curly brackets of (14) and the vector rand, respectively. Since each of these operations requires $O(N^6)$ elementary operations, the total computational complexity to compute a draw Π^m is $4 \times O(N^6)$. Also, computation of $\underline{\Omega}_\Pi^{-1}\text{vec}(\underline{\mu}_\Pi)$ requires $O(N^6)$ operations, but this is fixed across repetitions so it needs to be computed just once.[6]

For a system of 20 variables, the "medium" size in studies such as Banbura, Giannone, and Reichlin (2010), Carriero, Clark, and Marcellino (2016), Giannone, Lenza, and Primiceri (2015) and Koop (2013), this amounts to $4 \times 20^6 = 256$ million elementary operations (per single draw). This is the main bottleneck that so far prevented estimation of models with stochastic volatility using more than a handful of variables, typically 3 to 5.[7]

[5] In such case one needs to introduce another set of state variables s_T used to approximate the error term appearing in (4). In the case of volatilities independent across equations one could instead use the single-move sampler of Jacquier, Polson and Rossi (1994) and avoid drawing the mixture states s_T.

[6] Some speed improvements can be obtained as suggested by Chan (2015) by using

$$\text{vec}(\Pi^m) = C \backslash \left[C' \backslash \left\{ \text{vec}\left(\sum_{t=1}^{T} X_t y_t' \Sigma_t^{-1} \right) + \underline{\Omega}_\Pi^{-1} \text{vec}(\underline{\mu}_\Pi) \right\} + \text{rand} \right], \tag{15}$$

where C' is the Cholesky factor of $\bar{\Omega}_\Pi^{-1}$ and $\backslash$ stands for the command for backward solution of a linear system. While this is twice as fast as using (14), it is just a linear improvement and it is not sufficient to solve the bottleneck in estimation of large systems, as the overall computational complexity for calculating a draw is still of the order $O(N^6)$. In the remainder of the paper we use the strategy outlined in this footnote for all the models we consider.

[7] Carriero, Clark, and Marcellino (2016) estimate a larger system by assuming a specific structure for

2.3 Non-conjugate priors

The computational problem related to the dimension of the variance matrix of the coefficients can also arise in a homoskedastic setting. In particular, consider making the model (1)-(2) homoskedastic by substituting (2) with:

$$v_t \sim iid\ N(0, \Sigma). \tag{16}$$

For this model, commonly used prior distributions take a Normal-diffuse or independent Normal-Wishart form (e.g., Karlsson 2013). Although our results also apply in the Normal-diffuse case, we will focus on the independent Normal-Wishart (N-W) prior:

$$\begin{aligned} \text{vec}(\Pi) &\sim N(\text{vec}(\underline{\mu}_\Pi), \underline{\Omega}_\Pi); & (17)\\ \Sigma &\sim IW(\underline{d}_\Sigma \cdot \underline{\Sigma}, \underline{d}_\Sigma). & (18) \end{aligned}$$

The implied posteriors are

$$\begin{aligned} \text{vec}(\Pi)|\Sigma, y &\sim N(\text{vec}(\bar{\mu}_\Pi), \overline{\Omega}_\Pi); & (19)\\ \Sigma|\Pi, y &\sim IW((\underline{d}_\Sigma + T) \cdot \bar{\Sigma}, \underline{d}_\Sigma + T), & (20) \end{aligned}$$

with

$$\bar{\Omega}_\Pi^{-1} = \underline{\Omega}_\Pi^{-1} + \sum_{t=1}^{T} (\Sigma^{-1} \otimes X_t X_t'). \tag{21}$$

The matrix in (21) still has the same dimension as the one in (13), notwithstanding the fact that the matrix Σ does not vary with time.

The papers that have estimated homoskedastic VARs with a large cross section all use a different prior for Π, of the conjugate Normal-Wishart form:

$$\text{vec}(\Pi)|\Sigma \sim N(\text{vec}(\underline{\mu}_\Pi), \Sigma \otimes \Omega_0). \tag{22}$$

In this case, the prior is conditional on knowledge of Σ, and the matrix Σ is used to elicit the prior variance $\underline{\Omega}_\Pi = \Sigma \otimes \Omega_0$. Under these assumptions, the posterior variance becomes:

$$\bar{\Omega}_\Pi^{-1} = \Sigma^{-1} \otimes \left\{ \Omega_0^{-1} + \sum_{t=1}^{T} X_t X_t' \right\}, \tag{23}$$

---

the volatilities in the VAR, with Σ_t in (13) given by the product of a scalar σ_t and a constant matrix Σ ($\Sigma_t = \sigma_t \Sigma$), and with the prior variance $\underline{\Omega}_\Pi$ specified conditionally on the error variance, $\underline{\Omega}_\Pi = \Sigma \otimes \Omega_0$, where the Kronecker product constrains the prior to be symmetric across equations. Under these restrictions, equation (13) can be written as $\bar{\Omega}_\Pi^{-1} = \Sigma^{-1} \otimes \{\Omega_0^{-1} + \sum_{t=1}^{T} \sigma_t^{-1} X_t X_t'\}$, which does have a Kronecker structure and therefore can be easily handled. However, the assumption $\Sigma_t = \sigma_t \Sigma$ imposes a specific factor structure on the volatilities which implies that all the conditional volatilities are driven by a single factor (σ_t) with a loading of 1, and there is no idiosyncratic component. This setup implies that the order of magnitude of the movements in volatility is proportional across variables.

which has a Kronecker structure that permits manipulating the two terms in the Kronecker product separately (for details, see Carriero, Clark and Marcellino 2015), which provides huge computational gains and reduces the complexity to N^3. This specification allowed researchers, starting with Banbura, Giannone and Reichlin (2010), to estimate Bayesian VARs with more than a hundred variables.

However, a specification such as (22) is restrictive, as highlighted by Rothenberg (1963), Zellner (1973), Kadiyala and Karlsson (1993, 1997), and Sims and Zha (1998), and there are many situations in which the form (22) can turn out to be particularly unappealing.

First, it prevents permitting any asymmetry in the prior across equations, because the coefficients of each equation feature the same prior variance matrix Ω_0 (up to a scale factor given by the elements of Σ). For example, the traditional Minnesota prior in the original Litterman (1986) implementation cannot be cast in such a convenient form, because it imposes extra shrinkage on lags of variables that are not the lagged dependent variable in each equation. As another example, consider the case of a bivariate VAR in the variables y_1 and y_2 and suppose that the researcher has a strong prior belief that y_2 does not Granger cause y_1, while he/she has not strong beliefs that y_2 itself follows a univariate stationary process. This system of beliefs would require shrinking strongly towards zero the coefficients attached to y_2 in the equation for y_1. However, in order to keep the conjugate structure (22), this would also necessarily require shrinking strongly towards their prior means also the coefficients attached to y_2 in the equation for y_2, and this is unpleasant since the researcher does not have such strong priors in this respect.

Second, the Kronecker structure $\Sigma \otimes \Omega_0$ in (22) also implies the unappealing consequence that prior beliefs must be correlated across the equations of the reduced form representation of the VAR, with a correlation structure proportional to that of the disturbances (as described by the matrix Σ). Sims and Zha (1998) discuss this issue in depth, and propose an approach which allows for a more reasonable structure of the coefficient prior variance, and which also attains — like our proposal below — computational gains of order $O(N^2)$. Their approach is based on eliciting a prior featuring independence among the *structural* equations of the system, but does not achieve computational gains for an asymmetric prior on the *reduced form* equations' coefficients.[8]

[8] In particular, the approach of Sims and Zha (1998) achieves conceptual and computational gains by (i) working on the *structural* representation of the VAR, in which the matrix of the errors is diagonal (an identity matrix in their normalization scheme), and (ii) allowing independence across the coefficients belonging to different *structural* equations, which amounts to the prior variance of the coefficients being block-diagonal, which is desirable as it breaks the unreasonable symmetry across equations implied by the conjugate N-W prior. These two ingredients ensure that the posterior variance matrix has a block-diagonal structure, and

As we shall see, our estimation method solves the problems outlined above, making the independent N-W prior applicable in general, regardless of the size of the cross-section.

3 An estimation method for large VARs

We propose a very simple estimation method that solves the problems we discussed above by blocking the conditional posterior distribution in (9) in N different blocks.[9] Recall that in the step of the Gibbs sampler that involves drawing Π, all of the remaining model coefficients are given, and the matrix Σ_t is known. Defining $\Sigma_t^{0.5}$ as the lower-triangular Cholesky factor of Σ_t we have:[10]

$$\begin{bmatrix} v_{1,t} \\ v_{2,t} \\ \vdots \\ v_{N,t} \end{bmatrix} = \begin{bmatrix} \sigma^*_{1,1,t} & 0 & ... & 0 \\ \sigma^*_{2,1,t} & \sigma^*_{2,2,t} & & \vdots \\ \vdots & & \ddots & 0 \\ \sigma^*_{N,1,t} & ... & \sigma^*_{N,N-1,t} & \sigma^*_{N,N,t} \end{bmatrix} \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \vdots \\ \epsilon_{N,t} \end{bmatrix}, \tag{24}$$

where $\sigma^*_{j,i,t}$ denotes the generic (j,i)-th element of $\Sigma_t^{0.5}$. We will also denote by $\pi^{(j)}$ the vector of coefficients of equation j contained in column j of the matrix Π, for the intercept and coefficients on lagged y_t, and use $\pi^{(j)}_{i,l}$ to refer to the coefficient on lag l of variable i in

therefore achieves computational gains of order N^2. However, such a strategy still implies that the beliefs about the *reduced form* coefficients are correlated across equations in a way that depends on the covariance of the reduced form errors of the model, and gains are not attainable if one wants to impose an asymmetric prior on these *reduced form* coefficients, as explained in section 5.2 of their paper.

[9] Note that our triangularization applies to the draw of Π and as a result differs from the equation-by-equation approach used by Cogley and Sargent (2005) and Primiceri (2005) to draw the elements in the A matrix. In the case of the A matrix, the equation-by-equation approach obtains immediately from the posterior implied by the likelihood and prior. In the case of Π, the equation-by-equation approach requires the triangularization of the posterior associated with subtracting the appropriate linear combinations of residuals from both sides of the VAR's equations.

[10] That is, $\Sigma_t^{0.5}\Sigma_t^{0.5\prime} = \Sigma_t$. Of course, if a researcher is using the diagonalization in (3) then this matrix is readily available via $\Sigma_t^{0.5} = A^{-1}\Lambda_t^{0.5}$. However, it is important to stress that the triangularization illustrated here works for any error variance matrix Σ_t, not only those modeled as in (3).

equation j. The VAR can be written as:

$$\begin{aligned}
y_{1,t} &= \pi_0^{(1)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \pi_{i,l}^{(1)} y_{i,t-l} + \sigma_{1,1,t}^{*}\epsilon_{1,t} \\
y_{2,t} &= \pi_0^{(2)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \pi_{i,l}^{(2)} y_{i,t-l} + \sigma_{2,1,t}^{*}\epsilon_{1,t} + \sigma_{2,2,t}^{*}\epsilon_{2,t} \\
&\vdots \\
y_{N,t} &= \pi_0^{(N)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \pi_{i,l}^{(N)} y_{i,t-l} + \sigma_{N,1,t}^{*}\epsilon_{1,t} + \cdots + \sigma_{N,N-1,t}^{*}\epsilon_{N-1,t} + \sigma_{N,N,t}^{*}\epsilon_{N,t},
\end{aligned}$$

with the generic equation for variable j:

$$y_{j,t} - (\sigma_{j,1,t}^{*}\epsilon_{1,t} + \cdots + \sigma_{j,j-1,t}^{*}\epsilon_{j-1,t}) = \pi_0^{(j)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \pi_{i,l}^{(j)} y_{i,t-l} + \sigma_{j,j,t}^{*}\epsilon_{j,t}. \tag{25}$$

Consider estimating these equations in order from $j = 1$ to $j = N$. When estimating the generic equation j the term on the left hand side in (25) is known, since it is given by the difference between the dependent variable of that equation and the estimated residuals of all the previous $j-1$ equations. Therefore, we can define:

$$y_{j,t}^{*} = y_{j,t} - (\sigma_{j,1,t}^{*}\epsilon_{1,t} + \cdots + \sigma_{j,j-1,t}^{*}\epsilon_{j-1,t}), \tag{26}$$

and equation (25) becomes a standard generalized linear regression model for the variables in equation (26), with independent Gaussian disturbances with mean 0 and variance $\sigma_{j,j,t}^{*}$. The distribution (9) can be factorized as:

$$\begin{aligned}
p(\Pi|\Sigma_T, y) &= p(\pi^{(N)}|\pi^{(N-1)}, \pi^{(N-2)}, \ldots, \pi^{(1)}, \Sigma_T, y) \\
&\quad \times p(\pi^{(N-1)}|\pi^{(N-2)}, \ldots, \pi^{(1)}, \Sigma_T, y) \\
&\quad \vdots \\
&\quad \times p(\pi^{(1)}|\Sigma_T, y),
\end{aligned} \tag{27}$$

with generic element:

$$\begin{aligned}
p(\pi^{(j)}|\pi^{(j-1)}, \pi^{(j-2)}, \ldots, \pi^{(1)}, \Sigma_T, y) &= p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}}, \Sigma_T, y) \\
&\propto p(y|\Pi^{\{j\}}, \Pi^{\{1:j-1\}}, \Sigma_T) p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}}),
\end{aligned}$$

where $\Pi^{\{j\}} = \pi^{(j)}$ denotes the j-th column of the matrix Π and $\Pi^{\{1:j-1\}}$ all of the previous $1, ..., j-1$ columns. The term $p(y|\Pi^{\{j\}}, \Pi^{\{1:j-1\}}, A, \Lambda_T)$ is the likelihood of equation j, which coincides with the likelihood of the general linear regression model in (25). The term

$p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}})$ is the prior on the coefficients of the j-th equation, conditionally on the previous equations. The moments of $p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}})$ can be found recursively from the joint prior (6) using $p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}}) = p(\Pi^{\{j\}}, \Pi^{\{1:j-1\}})/p(\Pi^{\{1:j-1\}})$.

It follows that using the factorization in (27) together with the model in (25) allows one to draw the coefficients of the matrix Π in separate blocks $\Pi^{\{j\}}$ which can be obtained from:

$$\Pi^{\{j\}}|\Pi^{\{1:j-1\}}, \Sigma_T, y \sim N(\bar{\mu}_{\Pi^{\{j\}}}, \overline{\Omega}_{\Pi^{\{j\}}}), \tag{28}$$

with

$$\bar{\mu}_{\Pi^{\{j\}}} = \overline{\Omega}_{\Pi^{\{j\}}} \left\{ \underline{\Omega}^{-1}_{\Pi^{\{j\}}} \underline{\mu}_{\Pi^{\{j\}}} + \sum_{t=1}^{T} X_t \sigma^{*-1}_{j,j,t} y^*_{j,t} \right\}; \tag{29}$$

$$\overline{\Omega}^{-1}_{\Pi^{\{j\}}} = \underline{\Omega}^{-1}_{\Pi^{\{j\}}} + \sum_{t=1}^{T} X_t \sigma^{*-1}_{j,j,t} X_t', \tag{30}$$

where $y^*_{j,t}$ is defined in (26) and where $\underline{\Omega}^{-1}_{\Pi^{\{j\}}}$ and $\underline{\mu}_{\Pi^{\{j\}}}$ denote the prior moments on the j-th equation, given by the j-th column of $\underline{\mu}_{\Pi}$ and the j-th block on the diagonal of $\underline{\Omega}^{-1}_{\Pi}$. Note we have implicitly assumed here that $\underline{\Omega}^{-1}_{\Pi}$ is block diagonal, which means that we are ruling out any prior correlation among the coefficients belonging to different equations (i.e. $p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}}) = p(\Pi^{\{j\}})$). This assumption is frequent in the literature, but can be easily relaxed and we discuss how to do so below.[11] Therefore, the joint posterior distribution of Π can be simulated recursively in separate blocks $\Pi^{\{1\}}, \Pi^{\{2\}}|\Pi^{\{1\}}, \Pi^{\{3\}}|\Pi^{\{1:2\}}, ..., \Pi^{\{N\}}|\Pi^{\{1:N-1\}}$ using (28). Note that this amounts to simple Monte Carlo simulation which will produce draws numerically identical to those that would be obtained using system-wide estimation, meaning that any difference in the simulated posterior draws will be due to random variation (which eventually vanishes) and numerical rounding errors.

The dimension of the matrix $\overline{\Omega}^{-1}_{\Pi^{\{j\}}}$ in (30) is $(Np + 1)$, so that its manipulation only involves operations of order $O(N^3)$. However, since to obtain a draw for the full matrix Π one needs to draw separately all of its N columns, the total computational complexity of this estimation algorithm is $O(N^4)$, considerably smaller than the complexity of $O(N^6)$ implied by the standard estimation algorithm, with a gain of N^2. For a model with 20 variables this difference amounts to a 400-fold improvement in estimation time. Where is the computational gain coming from? In the traditional algorithm the sparsity implied by the possibility

[11] Some widely used priors within the independent N-W paradigm involve prior correlations among coefficients of the same equations, but not across equations. These include the sum of coefficients and unit root prior proposed by Sims (1993) and Sims and Zha (1998). As we already mentioned, the conjugate prior for a homoskedastic VAR in (22) does impose prior dependence across equations, but for this case an algorithm of computational complexity $O(N^3)$ is already available.

of triangularizing the system is not exploited, and all computations are carried out using the whole vectorized system. In our algorithm, instead, the triangularization allows one to estimate equations with $Np+1$ regressors, and the correlation among the different equations typical of SUR models is implicitly accounted for by the triangularization scheme.

While prior independence across equations is common in priors elicited in the literature, there might be cases in which a researcher wishes to specify priors which feature correlations across coefficients belonging to different equations. Examples include rational expectations, present-value models such as the expectation theory of the term structure of interest rates, the uncovered interest rate parity, and the permanent income hypothesis (see, e.g., Campbell and Shiller 1987). For this case, the general form of the posterior can be obtained easily using a similar triangularization argument on the joint prior distribution, and equation (28) generalizes to

$$\Pi^{\{j\}}|\Pi^{\{1:j-1\}}, \Sigma_T, y \sim N(\bar{\mu}_{\Pi^{\{j|1:j-1\}}}, \overline{\Omega}_{\Pi^{\{j|1:j-1\}}}), \tag{31}$$

with

$$\bar{\mu}_{\Pi^{\{j|1:j-1\}}} = \overline{\Omega}_{\Pi^{\{j|1:j-1\}}} \left\{ \underline{\Omega}^{-1}_{\Pi^{\{j|1:j-1\}}} \underline{\mu}_{\Pi^{\{j|1:j-1\}}} + \sum_{t=1}^{T} X_t \sigma^{*-1}_{j,j,t} y^*_{j,t} \right\}; \tag{32}$$

$$\overline{\Omega}^{-1}_{\Pi^{\{j|1:j-1\}}} = \underline{\Omega}^{-1}_{\Pi^{\{j|1:j-1\}}} + \sum_{t=1}^{T} X_t \sigma^{*-1}_{j,j,t} X_t', \tag{33}$$

where $\underline{\mu}_{\Pi^{\{j|1:j-1\}}}$ and $\underline{\Omega}_{\Pi^{\{j|1:j-1\}}}$ are the moments of $\Pi^{\{j\}}|\Pi^{\{1:j-1\}} \sim N(\underline{\mu}_{\Pi^{\{j|1:j-1\}}}, \underline{\Omega}_{\Pi^{\{j|1:j-1\}}})$, i.e. the conditional priors (for equation j conditional on all of the previous equations) implied by the joint prior specification. The conditional prior moments can be obtained recursively using (17) and standard results on multivariate Gaussian distributions:

$$\underline{\mu}_{\Pi^{\{j|1:j-1\}}} = \underline{\mu}_{\Pi^{\{j\}}} + \underline{\Omega}_{\Pi^{\{[j][1:j-1]\}}} \underline{\Omega}^{-1}_{\Pi^{\{[1:j-1][1:j-1]\}}} (\Pi^{\{1:j-1\}} - \underline{\mu}_{\Pi\{1:j-1\}}); \tag{34}$$

$$\underline{\Omega}_{\Pi^{\{j|1:j-1\}}} = \underline{\Omega}_{\Pi^{\{j\}}} - \underline{\Omega}_{\Pi^{\{[j][1:j-1]\}}} \underline{\Omega}^{-1}_{\Pi^{\{[1:j-1][1:j-1]\}}} \underline{\Omega}'_{\Pi^{\{[j][1:j-1]\}}}, \tag{35}$$

where $\underline{\Omega}_{\Pi^{\{j\}}}$ denotes the block of $\underline{\Omega}_{\Pi}$ corresponding to equation j, $\underline{\Omega}_{\Pi^{\{[1:j-1][1:j-1]\}}}$ all the blocks on the main block-diagonal, north-west of $\underline{\Omega}_{\Pi^{\{j\}}}$, and $\underline{\Omega}_{\Pi^{\{[j][1:j-1]\}}}$ all the blocks to the left of $\underline{\Omega}_{\Pi^{\{j\}}}$. The computational cost of deriving these conditional prior moments is negligible as they need to be computed only once outside the main MCMC sampler. Clearly in the case of a prior featuring independence across equations, $\underline{\Omega}_{\Pi^{\{[j][1:j-1]\}}}$ is a zero matrix and these expressions simplify to $\underline{\mu}_{\Pi^{\{j|1:j-1\}}} = \underline{\mu}_{\Pi^{\{j\}}}$ and $\underline{\Omega}_{\Pi^{\{j|1:j-1\}}} = \underline{\Omega}_{\Pi^{\{j\}}}$, yielding (29) and (30).

The non-conjugate priors common in the literature are entirely compatible with the algorithm described above, including the Minnesota prior (possibly with cross-variable shrinkage), the Sims and Zha (1998) priors (including the sum of coefficients and dummy initial

observation priors), the steady-state prior of Villani (2009), the long-run prior of Giannone, Lenza, and Primiceri (2016), and theory-based priors such as those of Ingram and Whiteman (1994) and Del Negro and Schorfheide (2004).

Finally, note that in a homoskedastic model the same reasoning for drawing the coefficients Π applies, so that the relevant posterior distributions for the Gibbs sampler would again be given by equation (28), with prior mean and variance given by (29) and (30) (or (31), (32), and (33) in case of prior dependence), with the only difference being that the subscript t would be omitted from the volatility terms $\sigma^*_{j,i,t}$. For this reason, the equation-by-equation step can be also used to estimate large VARs with asymmetric priors, such as, e.g., the Minnesota prior.[12]

3.1 The role of variable ordering

The fact that expression (24) and the following triangular system are based on a Cholesky decomposition of Σ_t might lead to the intuition that changing the ordering of the variables in the triangularization (24) would change the resulting draw of Π. However, this is not the case. The triangularization is used exclusively to obtain a draw from the conditional posterior Π by means of the recursion in (27), and the ordering of the equations within such recursion is completely inconsequential to the final result (i.e., the draw from $\Pi|\Sigma_T, y_T$).

Furthermore, it is worth clarifying that the Cholesky decomposition in (24) is simply used as an estimation device, not as a way to identify structural shocks. Once the MCMC simulations delivered posterior draws from the reduced form system, any identification scheme can be applied to perform structural analysis, including different Cholesky orderings, sign restrictions, and long run restrictions.

A more subtle point is related to the choice of the diagonalization (3) typically used in macroeconomics to model Σ_t. As noted by Sims and Zha (1998) and Primiceri (2005), since priors are elicited separately for the elements in the matrices A^{-1} and Λ_t, the implied prior of Σ_t is not invariant to the equation ordering. Clearly, different priors on Σ_t will lead to different posteriors, which means that different variable orderings would lead to different results. This problem — which we label the "prior ordering problem" — is not a feature of our algorithm, but rather it is inherent to all models using the diagonalization (3).

Models in which the prior is elicited directly on the matrix Σ_t do not suffer from the prior

[12] For the homoskedastic case Waggoner and Zha (2003) proposed an efficient Gibbs sampler also based on an equation-by-equation approach, and Koop, Korobilis, and Pettenuzzo (2016) proposed to use the method of compression to achieve computational gains. However these approaches are grounded on the Sims and Zha (1998) prior specification, and as such they cannot handle the case of asymmetric priors for the reduced form parameters.

ordering problem. The most prominent (and straightforward) example is the homoskedastic VAR with the non-conjugate prior given by (17) and (18), but of course this model features a constant error variance. Models featuring both a time-varying variance and a prior invariant to the ordering include Shin and Zhong (2016) (which uses the multivariate stochastic volatility specification of Philipov and Glickman 2006), and Bognanni (2018) (whose model has a reduced-form representation shared by all structural models in the class).

As mentioned above, the contribution of this paper is the triangularization of the draw from $\Pi|\Sigma_T, y_T$, and this is invariant to the ordering of the equations. However, since in our empirical application we follow the macroeconomics literature in modeling Σ_t via the diagonalization in (3), our empirical results are potentially affected by the prior ordering problem. The online Appendix (Section C) further discusses the prior ordering problem and shows that in our application it has a very mild effect on both the reduced form coefficients and the implied predictive densities.

4 A numerical comparison of the estimation methods

We now compare the proposed triangular algorithm with the traditional system-wide algorithm for estimation of the VAR-SV in (1)-(2). As detailed in the online Appendix (Section A), in this comparison and subsequent applications we use standard priors (independent Normal-Wishart for the VAR's coefficients, of the Minnesota prior form) and MCMC algorithms, but for the algorithm modifications associated with our triangularization.[13]

4.1 Computational complexity and speed of simulation

First, we compare the results obtained with the two algorithms as the dimension of the cross section N increases. We use monthly data taken from the dataset of McCracken and Ng (2016) (MN dataset), for the period January 1960 to December 2014, transformed as in their paper. Table 1 lists the 20 most widely followed, aggregate time series variables included in the medium-sized model used for many of the results reported below. The online Appendix (Section B) provides the full list of data series. We start by simply comparing computational times obtained using the two alternative algorithms, focusing on a medium-sized system of 20 variables and 13 lags. Of course, the two algorithms produce the same results. Importantly, though, the estimation of the model using the traditional system-wide algorithm was about 356 times slower. Our algorithm represents a substantial improvement in the ease of

[13] This means that we do impose cross-variable shrinkage, so the prior is asymmetric and could not be cast in the form (22).

estimating and handling these models, which is relevant especially in consideration of the fact that models of this size have been markedly supported by the empirical evidence.

Figure 1 illustrates the computational gains arising from the use of the triangular algorithm. The top panel shows the computational time (on a 3.5 GHz Intel Core i7) needed to perform 10 draws as a function of the size of the cross section using the triangular algorithm and the system-wide algorithm.[14] The bottom panel compares the gain in theoretical computational complexity (dashed line — which is equal to N^2) with the actual computational time. Since the computational gains become so large that they create scaling problems, results in this figure are displayed using a logarithmic vertical axis. As is clear, the computational gains from the triangular algorithm grow quadratically, and after $N = 25$ they become even larger than the theoretical gains, which we attribute to the fact that for such large systems the size of the operations is so large that it saturates the CPU computing power. Indeed, we do not extend this comparison to $N = 125$, which is the size used in the empirical application we present below in Section 5, because for a model of this size the system-wide algorithm would be extremely computationally demanding: a scalar number stored in double-precision floating-point format requires 8 bytes, and for a system with $N = 125$ the size of the covariance matrix of the coefficients is of dimension 203250, which would require about 330 GB of RAM ($203250^2 \times 8/10^9$).

4.2 Convergence and mixing

The traditional step-wise and our proposed triangular algorithm produce draws from the same posterior distribution. It could be argued that — as long as we have an increasing computing power — using the triangular algorithm only achieves gains in terms of speed. However, it is important to stress that — regardless of the power of the computers used to perform the simulation — the triangular algorithm will always produce many more draws than the traditional system-wide algorithm in the same unit of time. This has important consequences in terms of producing draws with good mixing and convergence properties.

To illustrate this point, we consider the quality of the draws that we can obtain from the two algorithms within a given amount of time. Specifically, for the 20-variable model with stochastic volatility described in the previous subsection, we first run the system-wide algorithm to obtain 5000 draws and record the total time needed to produce these draws. Then, we run our triangular algorithm for the same amount of time, and out of all the draws produced in this time interval, which are 356 times more — since our algorithm is about 356 times faster — we perform skip-sampling by saving only each 356-th draw.

[14] The size of the cross section is extended up to $N = 40$, using additional variables from the MN dataset.

Obviously, this exercise results in the same number of final draws (5000), but those obtained with our algorithm have dramatically improved convergence and mixing properties. **Figure 2** illustrates the recursive means for some selected coefficients and shows that the triangular algorithm with split-sampling reaches convergence much faster than the system-wide algorithm. This pattern is particularly marked for the volatility component of the model. In addition, inefficiency factors (see Figure A1 in the online Appendix) of 5000 draws obtained by running the two alternative algorithms *for the same amount of time* are much lower for draws produced by the triangular algorithm than for the system-wide algorithm. The triangular algorithm can produce in the same amount of time draws many times closer to i.i.d. sampling, which therefore feature better convergence properties. Instead, the system-wide algorithm is slower to converge (in a unit of time), especially so for the innovations to volatility and the volatility states.

Since these gains increase nonlinearly with the system size, we conclude that, for forecasting or structural analysis with medium and large Bayesian VARs, our estimation method based on the triangular algorithm offers computational gains large enough that many researchers should find it preferable. This should be especially true in forecasting analyses that involve model estimation at many different points in time.

5 A large structural VAR with drifting volatilities

In this section we summarize three key results of an application of structural analysis using our estimation method based on the triangular algorithm to estimate a very large VAR with stochastic volatility and asymmetric priors. The online Appendix provides the detailed results. In this application, we consider a VAR(13) with 125 variables, including all of the variables considered by McCracken and Ng (2016) with the exception of housing permits and their disaggregate components, which we exclude for their collinearity. The total number of objects to be estimated is extremely large: 203250 mean coefficients, 7750 covariance coefficients, 125 latent states (each of length T), and 7875 covariances of the states.

First, our algorithm makes feasible estimation of such a very large model, and works well. Despite the huge dimension of the system, our estimation algorithm can produce 5000 draws (after burning 500) in just above 7 hours on a 3.5 GHz Intel Core i7. Inefficiency factors and potential scale reduction factors for the various parameters and latent states indicate that, once a skip-sampling of 5 is performed — leaving 1000 clean draws — the convergence and mixing properties are good (see Figure C1 in the online Appendix (Section C.1)). Note that, with a model this large, skip-sampling greatly reduces storage costs.

Second, consistent with previous research, our estimates show considerable time variation in volatility (see Figures C2, C3, C4, and C5 in the online Appendix (Section C.1)). There is substantial homogeneity in the estimated volatility patterns for variables belonging to the same group, such as industrial production (IP) and producer price indexes (PPI) or interest rates at different maturities, but there is some heterogeneity across groups of variables. The Great Moderation starting around 1985 is much more evident when the data are aggregated to the quarterly frequency (Appendix Figures C4 and C5). The effects of the recent crisis are more heterogeneous. In particular, while volatility of real variables, such as IP and employment, and financial variables, such as stock price indexes, interest rates and spreads, goes back to lower levels after the peak associated with the crisis, there remains a much higher level of volatility than before the crisis in price indicators, in particular in the PPI and its components and in several CPI components, as well as in monetary aggregates and housing starts. Overall, the first principal component of all the estimated volatilities explains about 45% of overall variance, and the first three 73%, confirming that commonality is indeed present but idiosyncratic movements also matter (as in the GFSV specification of Carriero, Clark, and Marcellino 2016 and the factor volatility specification of Carriero, Clark, and Marcellino 2017).

Third, estimated impulse responses for a unitary shock to the federal funds rate (for identification, the federal funds rate is ordered after slow-moving and before fast-moving variables) display patterns in line with economic theory (see Figures C6 and C7 in the Appendix (Section C.1)). The estimates show a significant deterioration in real variables such as IP, unemployment, employment, and housing starts, only very limited evidence of a price puzzle, with most price responses not statistically significant, a significant deterioration in stock prices, a less than proportional increase in the entire term structure, which leads to a decrease in the term spreads, progressively diminishing over time, and a negative impact on the ISM indexes. Overall, the responses are in line with those reported in Banbura, Giannone and Reichlin (2010), since the presence of heteroskedasticity does not affect substantially the VAR coefficient estimates, but it matters for calculating the confidence bands and understanding the evolution of the size of the shock (and therefore of the actual responses that are proportional to the actual size of the shock) over time. Stochastic volatility would also matter for variance decompositions, omitted in the interest of brevity.

6 The role of model size and volatility for forecasting

To further illustrate the use of our proposed approach to large models, this section assesses the effects of time variation in volatility and a large information set on the accuracy of out-of-sample point and density forecasts from VARs.[15] The out-of-sample exercise is performed recursively, starting with the estimation sample 1960:3 to 1970:2 (ten years of monthly data) and ending with the estimation sample 1960:3 to 2014:5. We compute forecasts up to 12 step-ahead; therefore the forecasting samples range from 1970:3-1971:2 to 2014:6-2015:5, for a total of 531 sets of 12-step ahead forecasts.

We consider four models. The first model is a small homoskedastic VAR including the growth rate of industrial production ($\Delta \ln IP$), the inflation rate for the price index of consumption expenditures ($\Delta \ln PCEPI$) and the effective Federal Funds Rate (FEDFUNDS). The second model is also a homoskedastic VAR, but includes the 20 macroeconomic variables listed in Table 1. As similar models have been shown to be very competitive in forecasting in papers such as Banbura, Giannone, and Reichlin (2010), Carriero, Clark, and Marcellino (2015), Giannone, Lenza, and Primiceri (2015) and Koop (2013) we set this as our benchmark; namely, we will provide results relative to the performance of this model. The third model is still based on a tri-variate specification, but it allows for time variation in volatilities. Also, small models of this type have received support in the literature in terms of their forecasting performance; see, e.g., Clark (2011), Clark and Ravazzolo (2015), Cogley, Morozov, and Sargent (2005), and D'Agostino, Gambetti and Giannone (2013). Moreover, models of this scale have been used in the structural analyses of Cogley and Sargent (2005) and Primiceri (2005). The fourth model includes both time variation in the volatilities and a larger, medium-scale (20 variable) information set, thereby using both the ingredients that seem to be important to improve density and point forecasts. This model can be rather easily estimated using the approach proposed in this paper.

A priori, we expect the inclusion of time variation in volatilities to improve density forecasts via a better modeling of error variances, while the use of a larger dataset should improve point forecasts via a better specification of the conditional means. However, this is not the whole story, as there are also interaction effects: a better point forecast should improve the density forecast as well, by centering the predictive density around a more

[15] As noted by Diebold (2015), pseudo-out-of-sample forecasting exercises are not superior to several other model comparison techniques, notably F-tests and posterior odds, and are actually less powerful. However, performing posterior odds analysis presents problems in the case at hand because for the independent N-W prior used in this paper the marginal likelihood is not available in closed form and its computation would require an extremely demanding Monte Carlo integration.

reliable mean, and time-varying volatilities should improve the point forecasts — especially at longer horizons — because the heteroskedastic model will provide more efficient estimates (through a GLS argument) and therefore a better characterization of the predictive densities, with the predictive means gradually deviating from their homoskedastic counterparts as the predictive densities cumulate nonlinearly with the forecast horizon.

This is precisely the pattern we find in the data. The left panels of **Figure 3** display the Root Mean Squared Forecast Error (RMSFE) relative (ratio) to the benchmark (the 20-variable homoskedastic VAR), so that a value below 1 denotes a model outperforming the benchmark. The large homoskedastic model outperforms the small homoskedastic model for all variables at all horizons, suggesting that the inclusion of more data does improve the specification of the conditional means and therefore the point forecasts. The inclusion of time variation in volatilities consistently improves the performance of the small model, and for FEDFUNDS it also outperforms the benchmark at long horizons. However, the small heteroskedastic model is still largely dominated by the benchmark at short forecast horizons. The model with both time-varying volatilities and a large cross section instead provides systematically better point forecasts than the benchmark (and than the other models), with the only exception of inflation for the 1, 2, and 3 step-ahead horizons.

The right-hand panels of **Figure 3** present results for density forecasts, based on the average log scores. The figure displays the average log scores relative (difference) to the benchmark (the 20-variable homoskedastic VAR), so that a value above 0 denotes a model outperforming the benchmark. Both homoskedastic specifications perform quite poorly in density forecasting, while the heteroskedastic ones can achieve very high gains. Moreover, the large heteroskedastic system consistently outperforms the small heteroskedastic system. In combination with the findings for point forecasts, this result suggests that while both the heteroskedastic models provide a better assessment of the overall uncertainty around the forecasts, the model based on the large cross section centers such uncertainty around a more reliable mean, thereby obtaining further gains in predictive accuracy.

For the larger specifications (the VAR with 20 variables) it is of course possible to compare forecasts for all the variables included in the cross section. Results of this comparison are displayed in **Figure 4** (for point forecasts) and **Figure 5** (for density forecasts). In these graphs each subplot corresponds to a different variable.

In all of the subplots in **Figure 4** the x axes measure the RMSFE obtained by the large VAR when we allow for stochastic volatility, while the y axes measure the same loss function (RMSFE) obtained by the homoskedastic specification. Each point corresponds to a different forecast horizon, and when a point is *above* the 45 degree line this shows that

the RMSFE of the heteroskedastic specification is smaller, indicating that the inclusion of variation in the volatility improved point forecasting performance. As is clear in the graph, in several instances the models produce point forecasts of similar accuracy. However, as the forecast horizon increases (which can be indirectly inferred from the graph as in general higher RMSFE correspond to longer forecast horizons) the specification with variation in the volatilities tends to outperform the homoskedastic version of the model. The mechanism at play is as follows: the heteroskedastic model provides more efficient estimates and therefore a better characterization of the predictive densities, while the homoskedastic model is misspecified and therefore provides an inferior characterization of the predictive densities. At short forecast horizons this does not have much effect on point forecasts, but as the forecast horizon increases, the predictive densities cumulate nonlinearly and therefore the misspecification of the homoskedastic model increasingly reduces the relative accuracy.

We now turn to density forecasts, which are described in **Figure 5**. In the subplots in **Figure 5** the x axes measure the (log) density score obtained by the large VAR when we allow for stochastic volatility, while the y axes measure the same gain function (score) obtained by the homoskedastic specification. Each point corresponds to a different forecast horizon, and when a point is *below* the 45 degree line this shows that the score of the heteroskedastic specification is larger, indicating that the inclusion of variation in the volatility improved density forecasting performance. In **Figure 5** the improvement coming from the introduction of time variation in the volatilities is striking, and it is common to nearly all variables. Clearly, stochastic volatility improves the overall assessment of uncertainty with respect to the homoskedastic model, and it does so both directly, by simply using a better variance around the point estimates, and indirectly, by centering the densities towards improved point forecasts (as documented in **Figure 4**).

7 Conclusions

This paper introduced a new approach to estimation of large VARs with non-conjugate priors and drifting volatilities. The method is based on a straightforward triangularization of the system, and it is very simple to implement. Indeed, if a researcher already has algorithms to produce draws from a VAR with an independent N-W prior and stochastic volatility, only the step in which the conditional mean parameters are drawn needs to be modified, which can be easily done with a few lines of code.

The algorithm ensures computational gains of order N^2 with respect to the traditional algorithm used to estimate VARs with time-varying volatilities, and because of this it is

possible to achieve much better mixing and convergence properties compared to existing algorithms and substantial computational gains. This makes estimation of this type of model feasible regardless of the dimension of the system. Given its simplicity and the advantages in terms of speed, mixing, and convergence, we argue that the proposed algorithm should be preferred in empirical applications, especially those involving large datasets.

Moreover, our approach makes viable the estimation of models with independent N-W priors (as well as Normal-diffuse priors) of any model size. Since the independent N-W prior is much more flexible than the conjugate N-W prior, we argue that it should be preferred in most situations, including some in which the model is homoskedastic. The conjugate N-W prior imposes restrictions on the prior covariance matrix of the coefficients which can be in many instances undesirable, since it implies that the prior precision has to be the same (up to a scaling factor) in all equations, and that coefficients belonging to different equations have to be correlated, with a correlation structure proportional to that of the error variance.

We have illustrated the method by studying the effects of a monetary policy shock in a large VAR with stochastic volatilities. Finally, we have shown how, jointly, the inclusion of time-varying volatilities and the use of a large dataset improve point and density forecasts for macroeconomic and financial variables, with gains that are larger than what would be obtained by using these two ingredients separately.

In closing we want to highlight two caveats. First, while the independent N-W prior avoids putting on the data the straightjacket that the conjugate N-W does, the computation of the marginal likelihood is not as simple, while for the conjugate N-W prior it is available in closed form (for homoskedastic models). Second, while the model with stochastic volatility does produce dramatically superior density forecasts than its homoskedastic counterpart, some work is still needed to improve the density forecasts in the exact periods a large swing in volatilities takes place. Both these issues require further research.

References

[1] Bognanni, Mark, 2018. A Class of Time–Varying Parameter Structural VARs for Inference under Exact or Partial Identification, manuscript.

[2] Banbura, M., Giannone, D., and Reichlin, L., 2010. Large Bayesian vector autoregressions, Journal of Applied Econometrics 25, 71-92

[3] Bernanke, B., Boivin, J., and Eliasz, P., 2005. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach, Quarterly Journal of Economics 120, 387-422.

[4] Campbell, J., and Shiller, R., 1987. Cointegration and tests of present value models, Journal of Political Economy 95, 1062-1088.

[5] Carriero, A., Clark, T., and Marcellino, M., 2015. Bayesian VARs: specification choices and forecast accuracy, Journal of Applied Econometrics 30, 46-73.

[6] Carriero, A., Clark, T., and Marcellino, M., 2016. Common drifting volatility in large Bayesian VARs, Journal of Business and Economic Statistics 34, 375-390.

[7] Carriero, A., Clark, T., and Marcellino, M., 2017. Measuring uncertainty and its effects on the economy, Review of Economics and Statistics, forthcoming.

[8] Chan, J., 2015. Large Bayesian VARs: a flexible Kronecker error covariance structure, manuscript.

[9] Chib, S., and Greenberg, E., 1995. Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models, Journal of Econometrics 68, 339-360.

[10] Clark, T., 2011. Real-time density forecasts from BVARs with stochastic volatility, Journal of Business and Economic Statistics 29, 327-341.

[11] Clark, T., and Ravazzolo, F., 2015. Macroeconomic forecasting performance under alternative specifications of time-varying volatility, Journal of Applied Econometrics 30, 551-575.

[12] Cogley, T., Morozov, S., and Sargent, T., 2005. Bayesian fan charts for U.K. inflation: forecasting and sources of uncertainty in an evolving monetary system, Journal of Economic Dynamics and Control 29, 1893-1925.

[13] Cogley, T., and Sargent, T., 2005. Drifts and volatilities: monetary policies and outcomes in the post-WWII US, Review of Economic Dynamics 8, 262-302.

[14] D'Agostino, D., Gambetti, L., and Giannone, D., 2013. Macroeconomic forecasting and structural change, Journal of Applied Econometrics 28, 82-101.

[15] Del Negro, M., and Schorfheide, F., 2004. Priors from general equilibrium models for VARs, International Economic Review 45, 643-673.

[16] Diebold, F., 2015. Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold-Mariano tests, Journal of Business and Economic Statistics 33, 1-9.

[17] Geweke, J., and Whiteman, C., 2006. Bayesian forecasting, In: G. Elliott, C.W.J. Granger, and A. Timmermann, (Eds.), Handbook of Economic Forecasting, Volume 1, 3-80, Elsevier.

[18] Giannone, D., Lenza, M., and Primiceri, G., 2015. Prior selection for vector autoregressions, Review of Economics and Statistics 97, 436-451.

[19] Giannone, D., Lenza, M., and Primiceri, G., 2016. Priors for the long run, CEPR Discussion Paper No. DP11261.

[20] Ingram, B., and Whiteman, C., 1994. Supplanting the 'Minnesota' prior: forecasting macroeconomic time series using real business cycle model priors, Journal of Monetary Economics 34, 497-510.

[21] Jacquier, E., Polson, N., and Rossi, P., 2002. Bayesian analysis of stochastic volatility models, Journal of Business and Economic Statistics 20, 69-87.

[22] Kadiyala, K., and Karlsson, S., 1993. Forecasting with generalized Bayesian vector autoregressions, Journal of Forecasting 12, 365-378.

[23] Kadiyala, K., and Karlsson, S., 1997. Numerical methods for estimation and inference in Bayesian VAR models, Journal of Applied Econometrics 12, 99-132.

[24] Karlsson, S., 2013. Forecasting with Bayesian vector autoregression, In: G. Elliott and A. Timmermann, (Eds.), Handbook of Economic Forecasting, Volume 2, 791-897, Elsevier.

[25] Kim, S., Shephard, N. and Chib, S., 1998. Stochastic volatility: likelihood inference and comparison with ARCH models, Review of Economic Studies 65, 361-393.

[26] Koop, G., 2013. Forecasting with medium and large Bayesian VARs, Journal of Applied Econometrics 28, 177-203.

[27] Koop, G., and Korobilis, D., 2013. Large time-varying parameter VARs, Journal of Econometrics 177, 185-198.

[28] Koop, G., Korobilis, D., and Pettenuzzo, D., 2016. Bayesian compressed vector autoregressions, Journal of Econometrics, forthcoming.

[29] Korobilis, D., and Pettenuzzo, D., 2017. Adaptive Minnesota prior for high-dimensional vector autoregressions, manuscript, Brandeis University.

[30] Litterman, R., 1986. Forecasting with Bayesian vector autoregressions — five years of experience, Journal of Business and Economic Statistics 4, 25-38.

[31] McCracken, M., and Ng, S., 2016. FRED-MD: a monthly database for macroeconomic research, Journal of Business and Economic Statistics 34, 574-589.

[32] Philipov, A. and Glickman, M., 2006. Multivariate stochastic volatility via Wishart processes, Journal of Business and Economic Statistics 24, 313-328.

[33] Primiceri, G., 2005. time-varying structural vector autoregressions and monetary policy, Review of Economic Studies 72, 821-852.

[34] Rothenberg, T., 1963. A Bayesian analysis of simultaneous equation systems, report 6315, Econometric Institute, Netherlands School of Economics, Rotterdam.

[35] Shin, M., and Zhong, M., 2016. A new approach to identifying the real effects of uncertainty shocks, manuscript.

[36] Sims, C., 1993. A nine-variable probabilistic macroeconomic forecasting model, in J. Stock and M. Watson, (Eds.), Business Cycles, Indicators and Forecasting, University of Chicago Press, 179-212.

[37] Sims, C., and Zha, T., 1998. Bayesian methods for dynamic multivariate models, International Economic Review 39, 949-968.

[38] Villani, M., 2009. Steady-state priors for vector autoregressions, Journal of Applied Econometrics 24, 630-650.

[39] Waggoner, D., and Zha, T., 2003. A Gibbs sampler for structural vector autoregressions, Journal of Economic Dynamics and Control 28, 349-366.

[40] Zellner, A., 1973. An Introduction to Bayesian Inference in Econometrics, Wiley: New York.

Table 1: Variables in the 20-variable forecasting models

Variable	**Mnemonic**
Real personal income	RPI ($\Delta \ln$)
Real PCE	DPCERA3M086SBEA ($\Delta \ln$)
Real manufacturing and trade sales	CMRMTSPLx ($\Delta \ln$)
Industrial production	INDPRO ($\Delta \ln$)
Capacity utilization in manufacturing	CUMFNS
Civilian unemployment rate	UNRATE
Total nonfarm employment	PAYEMS ($\Delta \ln$)
Hours worked: goods-producing	CES0600000007 (ln)
Average hourly earnings: goods-producing	CES0600000008 ($\Delta \ln$)
PPI for finished goods	PPIFGS ($\Delta \ln$)
PPI for commodities	PPICMM ($\Delta \ln$)
PCE price index	PCEPI ($\Delta \ln$)
Federal funds rate	FEDFUNDS
Total housing starts	HOUST (ln)
S&P 500 price index	S&P 500 ($\Delta \ln$)
U.S.-U.K. exchange rate	EXUSUKx ($\Delta \ln$)
1 yr. Treasury - FEDFUNDS spread	T1YFFM
10 yr. Treasury - FEDFUNDS spread	T10YFFM
BAA - FEDFUNDS spread	BAAFFM
ISM: new orders index	NAPMNOI

Note: For those variables transformed for use in the forecasting models, the table indicates the transformation in parentheses following the variable description.

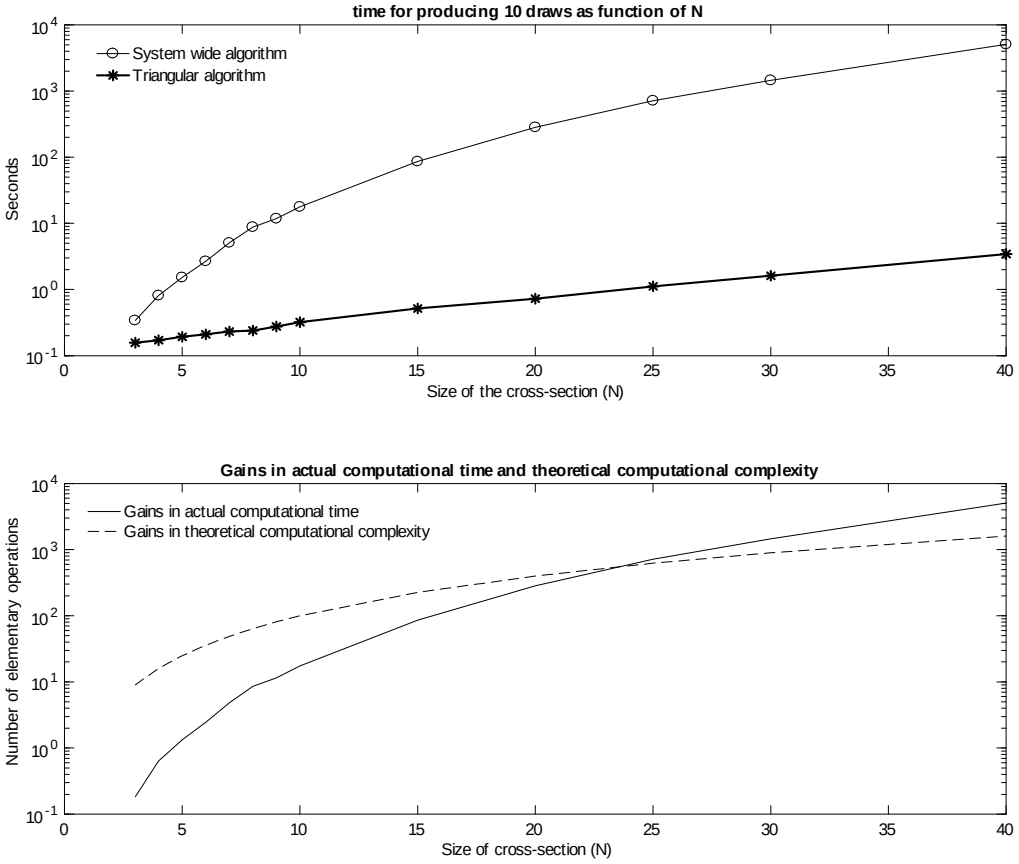

Figure 1: Actual computational time and theoretical computational complexity of the alternative algorithms. Note that due to the exponential nature of the gains the y-axes are in logarithmic scale. Computational times are computed as the average time (over 10 independent chains) required to draw 10 draws on a 3.5 GHz Intel Core i7.

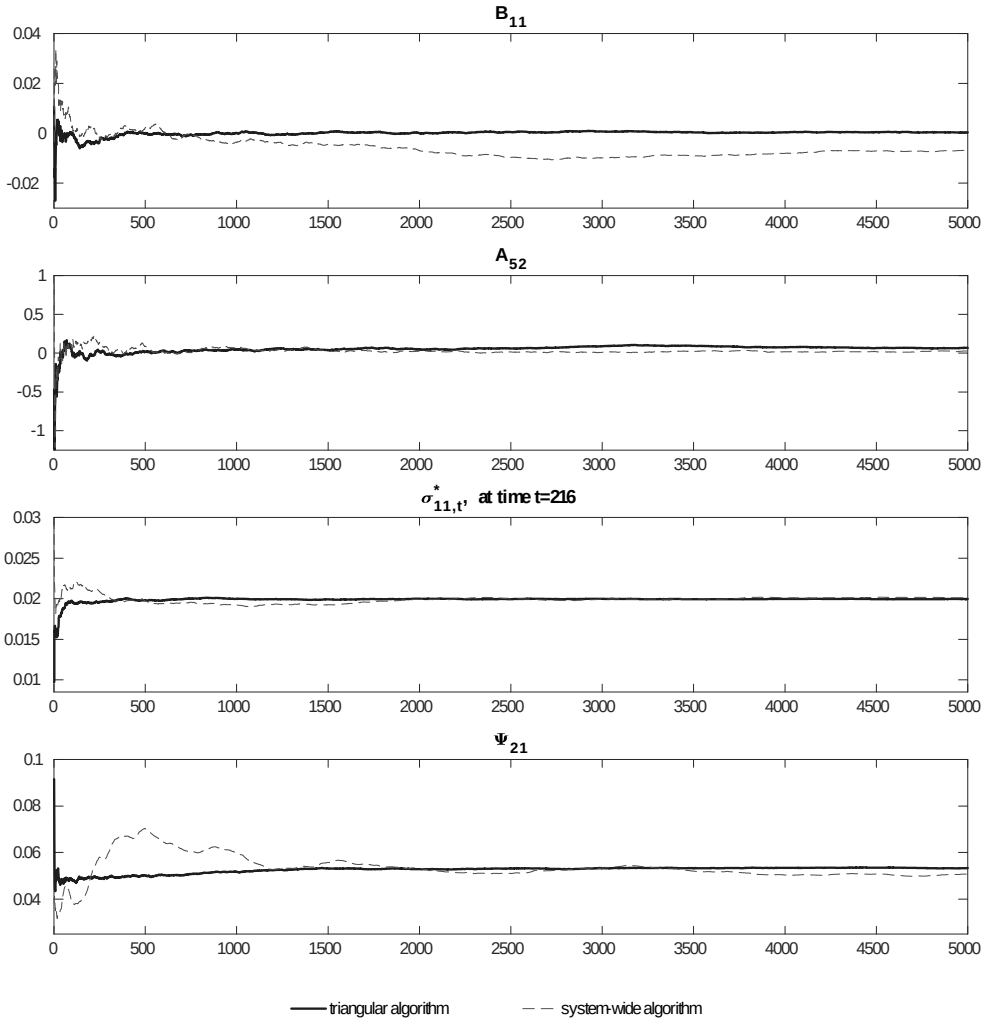

Figure 2: Recursive means of selected coefficients. Comparison between the system wide and triangular algorithm. The chains are initialised at the same value (set equal to the priors).

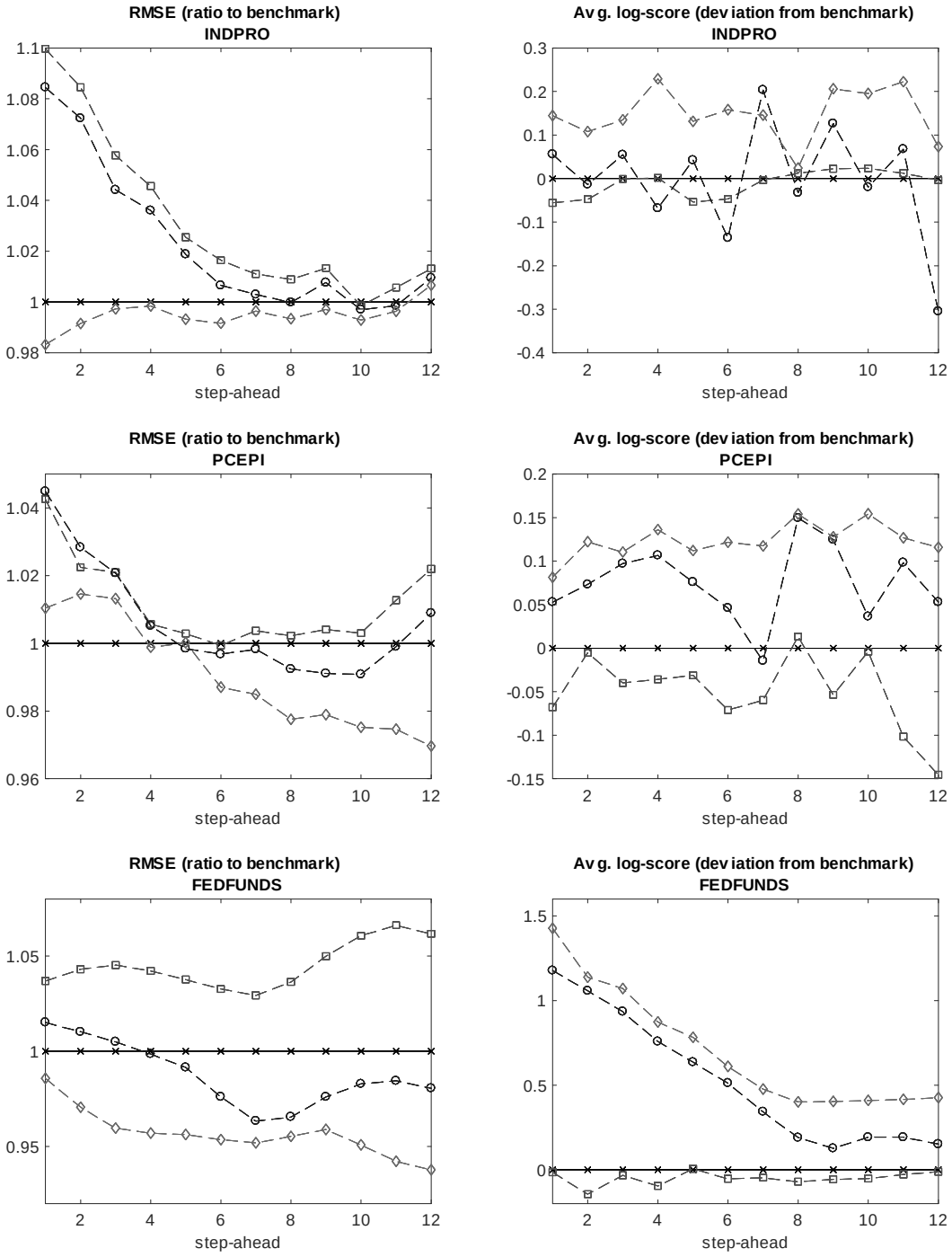

Figure 3: Forecast comparisons. Panels on the left hand side contain results for the point forecasts (relative RMSE of different models vs benchmark). The panels on the right hand side contain results for the density forecasts (Log-score gains of different models vs benchmark). In all the panels crosses represent the homoskedastic VAR with 20 variables (the benchmark model), the squares represent the homoskedastic VAR with 3 variables, the circles represent the heteroskedastic VAR with 3 variables, and the diamonds represent the heteroskedastic VAR with 20 variables.

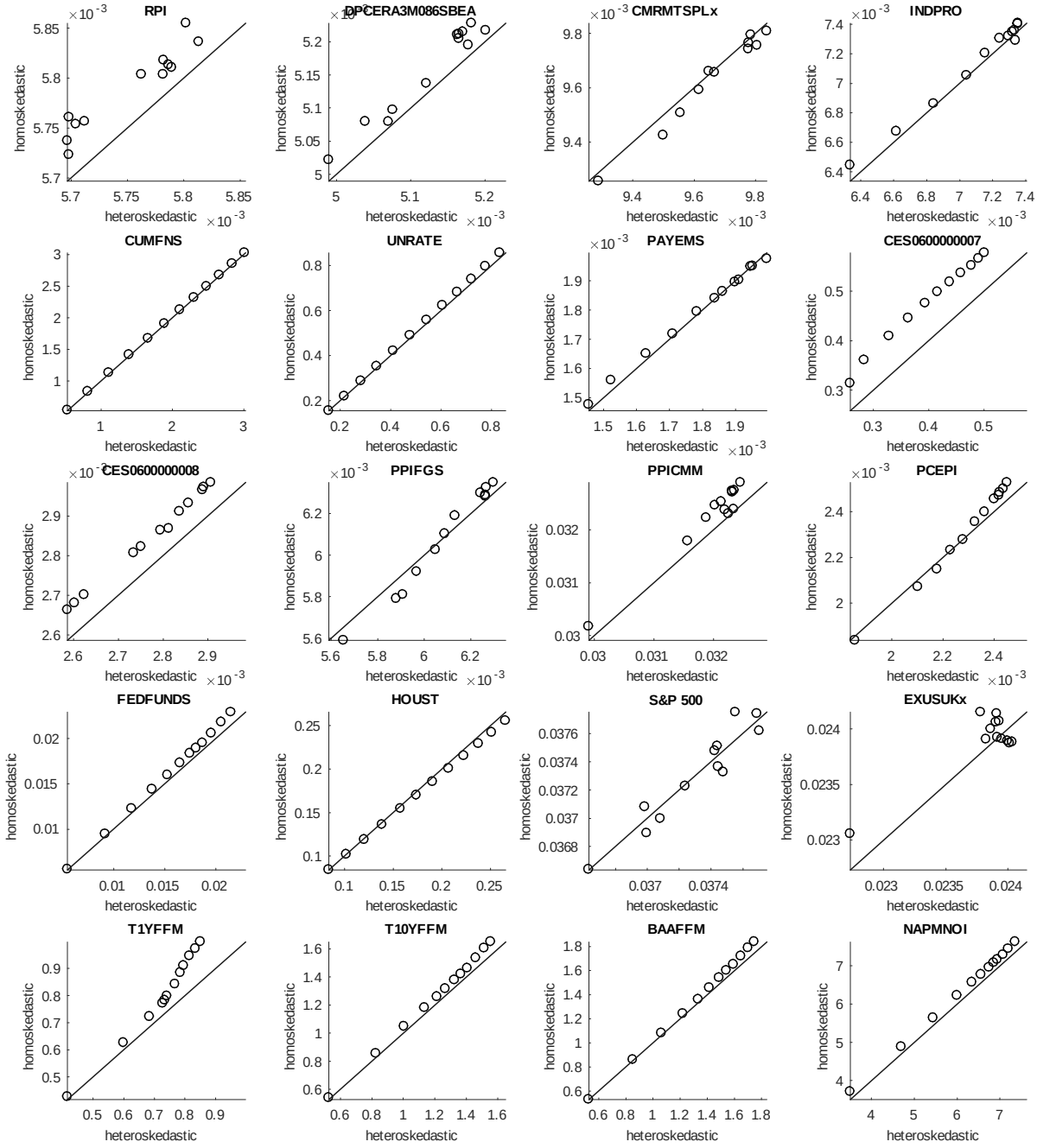

Figure 4: Comparison of point forecast accuracy. Each panel describes a different variable. The x axis reports the RMSFE obtained using the BVAR with stochastic volatility (heteroskedastic), the y axis reports the RMSFE obtained using the homoskedastic BVAR. Each point corresponds to a different forecast horizon from 1 to 12 step-ahead (in most cases, a higher RMSFE corresponds to a longer forecast horizon).

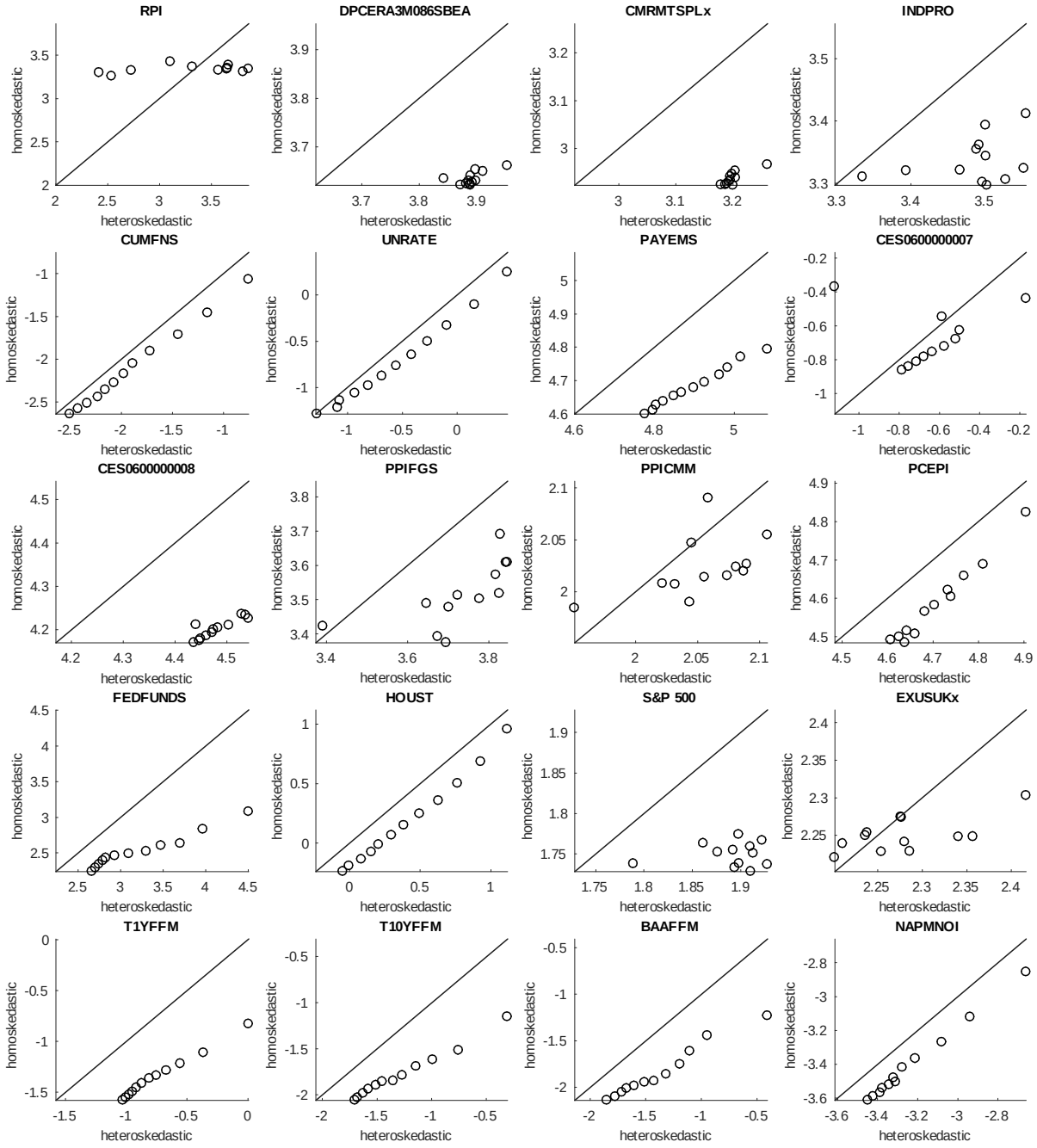

Figure 5: Comparison of density forecast accuracy. Each panel describes a different variable. The x axis reports the (log) density score obtained using the BVAR with stochastic volatility (heteroskedastic), the y axis reports the (log) density score obtained using the homoskedastic BVAR. Each point corresponds to a different forecast horizon from 1 to 12 step-ahead.